Résumés de thèses

Rubrique préparée par Sylvain Pogodalla

INRIA, Villers-lès-Nancy, F-54600, France
Université de Lorraine, LORIA, UMR 7503, Vandœuvre-lès-Nancy, F-54500, France
CNRS, LORIA, UMR 7503, Vandœuvre-lès-Nancy, F-54500, France
sylvain.pogodalla@inria.fr

Jean-Philippe FAUCONNIER : jean.philippe.fauconnier@gmail.com

Titre : Acquisition de liens sémantiques à partir d'éléments de mise en forme des textes : exploitation des structures énumératives

Mots-clés : Extraction de relations, analyse du document, mise en forme, apprentissage artificiel, discours.

Title: *Acquisition of Semantic Relations from Layout Elements: Exploitation of Enumerative Structures*

Keywords: *Relation extraction, document analysis, text layout, machine learning, discourse.*

Thèse de doctorat en Informatique, Institut de Recherche en Informatique de Toulouse, IRIT, UMR 5055, Université Toulouse III – Paul Sabatier, sous la direction de Nathalie Aussenac-Gilles (DR, CNRS, IRIT, UMR 5055, Toulouse) et Mouna Kamel (MC, Université de Perpignan). Thèse soutenue le 27/01/2016.

Jury : Mme Nathalie Aussenac-Gilles (DR, CNRS, IRIT, UMR 5055, Toulouse, codirectrice), Mme Mouna Kamel (MC, Université de Perpignan, codirectrice), M. Thierry Poibeau (DR, CNRS, LaTTiCe, rapporteur), Mme Pascale Sébillot (Pr, INSA de Rennes, IRISA, rapporteur), Mme Béatrice Daille (Pr, Université de Nantes, LINA, présidente), M. Olivier Ferret (IR HDR, CEA LIST, LVIC, examinateur), Mme Núria Gala (MC HDR, Université d'Aix-Marseille, LIF, examinatrice).

Résumé : *Ces dernières années de nombreux progrès ont été faits dans le domaine de l'extraction de relations à partir de textes, facilitant ainsi la construction de ressources lexicales ou sémantiques. Cependant, les méthodes proposées (apprentissage*

supervisé, méthodes à noyaux, apprentissage distant, etc.) n'exploitent pas tout le potentiel des textes : elles ont généralement été appliquées à un niveau phrastique, sans tenir compte des éléments de mise en forme.

Dans ce contexte, l'objectif de cette thèse est d'adapter ces méthodes à l'extraction de relations exprimées au-delà des frontières de la phrase. Pour cela, nous nous appuyons sur la sémantique véhiculée par les indices typographiques (puces, emphases, etc.) et dispositionnels (indentations visuelles, retours à la ligne, etc.) qui complètent des formulations strictement discursives. En particulier, nous étudions les structures énumératives verticales qui, bien qu'affichant des discontinuités entre leurs différents composants, présentent un tout sur le plan sémantique. Ces structures textuelles sont souvent révélatrices de relations hiérarchiques.

Notre travail est divisé en deux parties. La première partie décrit un modèle pour représenter la structure hiérarchique des documents. Ce modèle se positionne dans la suite des modèles théoriques proposés pour rendre compte de l'architecture textuelle : une abstraction de la mise en forme et une connexion forte avec la structure rhétorique sont faites. Toutefois, notre modèle se démarque par une perspective d'analyse automatique des textes. Nous en proposons une implémentation efficace sous la forme d'une méthode ascendante et nous l'évaluons sur un corpus de documents PDF.

La seconde partie porte sur l'intégration de ce modèle dans le processus d'extraction de relations. Plus particulièrement, nous nous sommes focalisés sur les structures énumératives verticales. Un corpus a été annoté selon une typologie multi-dimensionnelle permettant de caractériser et de cibler les structures énumératives verticales porteuses de relations utiles à la création de ressources. Les observations faites en corpus ont conduit à procéder en deux étapes par apprentissage supervisé pour analyser ces structures : qualifier la relation puis en extraire les arguments. L'évaluation de cette méthode montre que l'exploitation de la mise en forme, combinée à un faisceau d'indices lexico-syntaxiques, améliore les résultats.

URL où le mémoire pourra être téléchargé :
https://tel.archives-ouvertes.fr/tel-01324765

Pierre MARCHAL : pierre.inalco@gmail.com

Titre : Acquisition de schémas prédicatifs verbaux en japonais

Mots-clés : Japonais, linguistique de corpus, syntaxe, verbe.

Title: *Verbal Predicate-Frame Acquisition in Japanese*

Keywords: *Japanese, corpus linguistics, syntax, verb.*

Thèse de doctorat en Traitement automatique des langues, Institut National des Langues et Civilisations Orientales — INaLCO —, Paris, sous la direction de Thierry Poibeau (DR, CNRS, LaTTiCe). Thèse soutenue le 15/10/2015.

Jury : M. Thierry Poibeau (DR, CNRS, LaTTiCe, directeur), M. Raoul Blin (CR, CNRS, Centre de recherches linguistiques sur l'Asie orientale — CRLAO —, Paris, examinateur), Mme Cécile Fabre (Pr, Université de Toulouse II – Jean Jaurès, rapporteur et présidente), M. Kyô Kageura (Pr, Université de Tôkyô, Japon, rapporteur), M. Yves Lepage (Pr, Université Waseda, Japon, examinateur), Mme Frédérique Segond (Pr associé, INaLCO, Paris, examinatrice).

Résumé : *L'acquisition de connaissances relatives aux constructions verbales est une question importante pour le traitement automatique des langues, mais aussi pour la lexicographie qui vise à documenter les nouveaux usages linguistiques. Cette tâche pose de nombreux enjeux, techniques et théoriques. Dans le cadre de cette thèse, nous nous intéressons plus particulièrement à deux aspects fondamentaux de la description du verbe : la notion d'entrée lexicale et la distinction entre arguments et circonstants.*

À la suite de précédentes études en traitement automatique des langues et en linguistique nous posons qu'il existe un continuum entre homonymes et monosèmes ; de même nous faisons l'hypothèse qu'il n'y a pas de distinction marquée entre arguments et circonstants. Nous proposons une chaîne de traitement complète pour l'acquisition de schémas prédicatifs verbaux en japonais à partir d'un corpus non étiqueté de textes journalistiques. Cette chaîne de traitement intègre la notion d'argumentalité au processus de création des entrées lexicales et met en œuvre une modélisation de ces deux continuums. La ressource produite a fait l'objet d'une évaluation comparative qualitative, qui a permis de mettre en évidence la difficulté des ressources linguistiques à décrire de nouvelles données, plaidant par là même pour une lexicologie s'inscrivant dans le cadre épistémologique de la linguistique de corpus.

URL où le mémoire pourra être téléchargé :
http://www.theses.fr/2015INAL0015

Corentin RIBEYRE : corentin.ribeyre@gmail.com

Titre : Méthodes d'analyse supervisée pour l'interface syntaxe-sémantique : de la réécriture de graphes à l'analyse par transitions

Mots-clés : Interface syntaxe-sémantique, syntaxe profonde, analyse par transitions, graphes, réécriture de graphes.

Title: *Data-driven Methods for Syntax-Semantic Interface: from Graph Rewriting to Transition-based Parsing*

Keywords: *Syntax-semantic interface, deep syntax, transition-based parsing, graphs, graph rewriting.*

Thèse de doctorat en Sciences du Langage, Alpage, UMRI-001, Linguistique, Université Paris Diderot –Paris 7, sous la direction de Laurence Danlos (Pr, Université Paris Diderot – Paris 7), Djamé Seddah (MC, Université Paris-Sorbonne – Paris 4) et

Éric Villemonte de la Clergerie (CR, INRIA Paris – Rocquencourt). Thèse soutenue le 27/01/2016.

Jury : Mme Laurence Danlos (Pr, Université Paris Diderot – Paris 7, codirectrice), Mme Paola Merlo (Pr, Université de Genève, Suisse, rapporteur), M. John A. Carroll (Pr, University of Sussex, Royaume-Uni, rapporteur), M. Sylvain Kahane (Pr, Université Paris Ouest Nanterre La Défense, président), M. Djamé Seddah (MC, Université Paris-Sorbonne – Paris 4, codirecteur), M. Éric Villemonte de la Clergerie (CR, INRIA Paris – Rocquencourt, codirecteur).

Résumé : *Aujourd'hui, le volume de données textuelles disponibles est colossal. Ces données représentent une manne d'informations inestimables qu'il n'est pas possible de traiter manuellement. De fait, il est essentiel d'utiliser des techniques de Traitement Automatique des Langues pour arriver à extraire les informations saillantes et comprendre le sens sous-jacent. Cette thèse s'inscrit donc dans cette perspective et propose des ressources, des modèles et des méthodes pour permettre : (i) l'annotation automatique de corpus à l'interface entre la syntaxe et la sémantique afin d'en extraire la structure argumentale liant les prédicats (verbaux) à leurs arguments ; (ii) l'exploitation de ces ressources grâce à des méthodes efficaces.*

Dans un premier temps, nous proposons un système de réécriture de graphes, ainsi qu'un ensemble de règles de réécriture manuellement écrites permettant l'annotation automatique de la syntaxe profonde du français. Grâce à cette approche, deux corpus ont vu le jour, à savoir le DeepSequoia*, une version en syntaxe profonde du corpus* Séquoia*, et le* DeepFTB*, une version en syntaxe profonde du* French Treebank *en dépendances.*

Dans un second temps, nous proposons deux extensions d'analyseurs par transitions et les adaptons à l'analyse de graphes. Nous développons également un ensemble de traits linguistiquement riches, issus d'analyses syntaxiques. L'idée est d'apporter des informations topologiquement diversifiées donnant à nos analyseurs les indices nécessaires pour une prédiction performante de la structure argumentale. Couplé enfin à un analyseur par factorisation d'arcs, cet ensemble de traits permet d'établir l'état de l'art sur le français et de dépasser celui préalablement établi pour les corpus DM *(corpus de dépendences, dérivées des MRS de* Deepbank*) et* Enju's Predicate Argument Structure *disponibles pour l'anglais.*

Enfin, nous explorons succinctement une méthode d'induction automatisant le passage d'une représentation surfacique (un arbre) vers une représentation sémantique (un graphe) en utilisant des techniques de fouille de sous-graphes fréquents. Cette méthode complète alors notre ensemble cohérent de ressources et modèles proposés pour l'analyse de l'interface syntaxe-sémantique sur les langues française et anglaise.

URL où le mémoire pourra être téléchargé :
https://hal.archives-ouvertes.fr/tel-01323245